Exploratory Data Analysis

Lecture 10

Corina Besliu

Technical University of Moldova

October 3, 2021

Recap Hypothesis Testing

1. Formulate the hypothesis
   $H_0 : \beta_k = 0$
   $H_A : \beta_k \neq 0$

Recap Hypothesis Testing

1. Formulate the hypothesis
   $H_0 : \beta_k = 0$
   $H_A : \beta_k \neq 0$
2. Compute the t-statistic:

$$t_k = \frac{(\hat{\beta}_k - \beta_{H_0})}{SE(\hat{\beta}_k)} \qquad (k = 1, 2, \ldots, K)$$

3. Find t-critical in the t-table for chosen significance level α (commonly $\alpha = 0.05$) and respective degrees of freedom (df = N - K - 1)
4. Reject H_0 if $-t_c < t_k > t_c$

Recap Hypothesis Testing

1. Formulate the hypothesis
   $H_0 : \beta_k = 0$
   $H_A : \beta_k \neq 0$
2. Compute the t-statistic:

$$t_k = \frac{(\hat{\beta}_k - \beta_{H_0})}{SE(\hat{\beta}_k)} \qquad (k = 1, 2, \ldots, K)$$

3. Find t-critical in the t-table for chosen significance level α (commonly $\alpha = 0.05$) and respective degrees of freedom (df = N - K - 1)
4. Reject H_0 if $-t_c < t_k > t_c$

Recap Hypothesis Testing

1. Formulate the hypothesis
   $H_0 : \beta_k = 0$
   $H_A : \beta_k \neq 0$
2. Compute the t-statistic:

$$t_k = \frac{(\hat{\beta}_k - \beta_{H_0})}{SE(\hat{\beta}_k)} \qquad (k = 1, 2, \ldots, K)$$

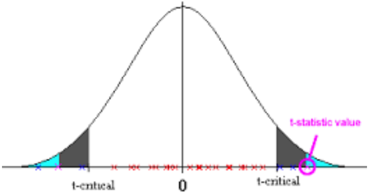
3. Find t-critical in the t-table for chosen significance level α (commonly $\alpha = 0.05$) and respective degrees of freedom (df = N - K - 1)
4. Reject H_0 if $-t_c < t_k > t_c$

Recap Hypothesis Testing

1. Formulate the hypothesis
   $H_0 : \beta_k = 0$
   $H_A : \beta_k \neq 0$
2. Compute the t-statistic:

$$t_k = \frac{(\hat{\beta}_k - \beta_{H_0})}{SE(\hat{\beta}_k)} \qquad (k = 1, 2, \ldots, K)$$

3. Find t-critical in the t-table for chosen significance level α (commonly $\alpha = 0.05$) and respective degrees of freedom (df = N - K - 1)
4. Reject H_0 if $-t_c < t_k > t_c$



p-values

A p-value for a t_k is the probability of observing a t-score that size or larger (in absolute value) if H_0 were true ($0 \geq p - value \leq 1$).

In other words:

p-value is the probability that you rejected a H_0 that was correct (assuming that the estimate is in the expected direction).

Graphically, it's two times the area under the curve of the t-distribution between $|t_k|$ and ∞.

p-values

A p-value for a t_k is the probability of observing a t-score that size or larger (in absolute value) if H_0 were true $(0 \geq p - value \leq 1)$.

In other words:
p-value is the probability that you rejected a H_0 that was correct (assuming that the estimate is in the expected direction).

Graphically, it's two times the area under the curve of the t-distribution between $|t_k|$ and ∞.

p-values

A p-value for a t_k is the probability of observing a t-score that size or larger (in absolute value) if H_0 were true ($0 \geq p - value \leq 1$).

In other words:
p-value is the probability that you rejected a H_0 that was correct (assuming that the estimate is in the expected direction).

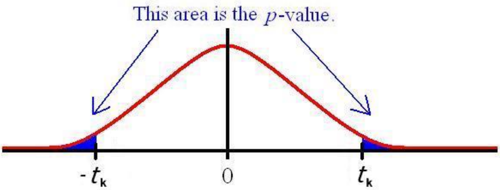
Graphically, it's two times the area under the curve of the t-distribution between $|t_k|$ and ∞.

p-values

A p-value for a t_k is the probability of observing a t-score that size or larger (in absolute value) if H_0 were true ($0 \geq p - value \leq 1$).

<u>In other words:</u>
p-value is the probability that you rejected a H_0 that was correct (assuming that the estimate is in the expected direction).

Graphically, it's two times the area under the curve of the t-distribution between $|t_k|$ and ∞.



p-value Rule of Thumb

To cofidently reject a null, you will want a low p-value

Rule of Thumb:

$p - value < 0.05 \rightarrow$ reject H_0

If $p - value > 0.05 \rightarrow \alpha > 0.05 \rightarrow \underbrace{\text{P(Type I Error)} > 0.05}_{\text{Not Great!}}$

p-value Rule of Thumb

To cofidently reject a null, you will want a low p-value

Rule of Thumb:

$$p - value < 0.05 \rightarrow \text{reject } H_0$$

If $p - value > 0.05 \rightarrow \alpha > 0.05 \rightarrow \underbrace{\text{P(Type I Error)} > 0.05}_{\text{Not Great!}}$

p-value Rule of Thumb

To cofidently reject a null, you will want a low p-value

Rule of Thumb:

$$p - value < 0.05 \rightarrow \text{reject } H_0$$

If $p - value > 0.05 \rightarrow \alpha > 0.05 \rightarrow \underbrace{\text{P(Type I Error)} > 0.05}_{\text{Not Great!}}$

p-value Rule of Thumb

To cofidently reject a null, you will want a low p-value

Rule of Thumb:

$$p - value < 0.05 \rightarrow \text{reject } H_0$$

If $p - value > 0.05 \rightarrow \alpha > 0.05 \rightarrow \underbrace{\text{P(Type I Error)} > 0.05}_{\text{Not Great!}}$

p-value Rule of Thumb

To cofidently reject a null, you will want a low p-value

Rule of Thumb:

$$p - value < 0.05 \rightarrow \text{reject } H_0$$

If $p - value > 0.05 \rightarrow \alpha > 0.05 \rightarrow \underbrace{\text{P(Type I Error)} > 0.05}_{\text{Not Great!}}$

p-value Rule of Thumb

To cofidently reject a null, you will want a low p-value

Rule of Thumb:

$$p - value < 0.05 \rightarrow \text{reject } H_0$$

$$\text{If } p - value > 0.05 \rightarrow \alpha > 0.05 \rightarrow \underbrace{\text{P(Type I Error)} > 0.05}_{\text{Not Great!}}$$

Regression Results

```
                            OLS Regression Results
=======================================================================================
Dep. Variable:                 Profit   R-squared (uncentered):                  0.984
Model:                            OLS   Adj. R-squared (uncentered):             0.982
Method:                 Least Squares   F-statistic:                             839.8
Date:                Wed, 29 Sep 2021   Prob (F-statistic):                   2.95e-49
Time:                        11:27:08   Log-Likelihood:                        -661.85
No. Observations:                  60   AIC:                                     1332.
Df Residuals:                      56   BIC:                                     1340.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
RD             0.7145      0.068     10.490      0.000       0.578       0.851
Admin          0.2546      0.043      5.927      0.000       0.169       0.341
Marketing      0.0962      0.025      3.915      0.000       0.047       0.145
Office         0.9425      0.663      1.421      0.161      -0.386       2.271
==============================================================================
Omnibus:                        6.429   Durbin-Watson:                   1.745
Prob(Omnibus):                  0.040   Jarque-Bera (JB):                9.643
Skew:                          -0.196   Prob(JB):                      0.00805
Kurtosis:                       4.924   Cond. No.                         91.1
==============================================================================
```

F-test

Although R^2 and $\bar{R}^2$ measure the overall degree of fit of an equation, they don't provide a formal hypothesis test of that overall fit.

But we have the F-test!

The F-test is a joint hypothesis test.

$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$
$H_A : H_0$ in not true

F-test

Although R^2 and $\bar{R}^2$ measure the overall degree of fit of an equation, they don't provide a formal hypothesis test of that overall fit.

But we have the F-test!

The F-test is a joint hypothesis test.

$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$
$H_A : H_0$ in not true

F-test

Although R^2 and $\bar{R}^2$ measure the overall degree of fit of an equation, they don't provide a formal hypothesis test of that overall fit.

But we have the F-test!

The F-test is a joint hypothesis test.

$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$
$H_A : H_0$ in not true

F-test

Although R^2 and $\bar{R}^2$ measure the overall degree of fit of an equation, they don't provide a formal hypothesis test of that overall fit.

But we have the F-test!

The F-test is a joint hypothesis test.

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$H_A : H_0 \text{ in not true}$$

F-test Steps

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$H_A : H_0 \text{ in not true}$$

1. Translate the null hypothesis into constraints on the equation. $\rightarrow$
   Constrained equation: $Y_i = \beta_0 + \epsilon_i$
   Unconstrained equation: $Y_i = \beta_0 + \beta_1 + \beta_2 + ... + \beta_k + \epsilon_i$

F-test Steps

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$H_A : H_0 \text{ in not true}$$

1. Translate the null hypothesis into constraints on the equation. $\rightarrow$
   Constrained equation: $Y_i = \beta_0 + \epsilon_i$
   Unconstrained equation: $Y_i = \beta_0 + \beta_1 + \beta_2 + ... + \beta_k + \epsilon_i$

F-test Steps

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$H_A : H_0 \text{ in not true}$$

1. Translate the null hypothesis into constraints on the equation. $\rightarrow$
   Constrained equation: $Y_i = \beta_0 + \epsilon_i$
   Unconstrained equation: $Y_i = \beta_0 + \beta_1 + \beta_2 + ... + \beta_k + \epsilon_i$

F-test Steps

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$H_A : H_0 \text{ in not true}$$

1. Translate the null hypothesis into constraints on the equation. $\rightarrow$
   Constrained equation: $Y_i = \beta_0 + \epsilon_i$
   Unconstrained equation: $Y_i = \beta_0 + \beta_1 + \beta_2 + ... + \beta_k + \epsilon_i$

F-test Steps

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$H_A : H_0 \text{ in not true}$$

1. Translate the null hypothesis into constraints on the equation. $\rightarrow$
   Constrained equation: $Y_i = \beta_0 + \epsilon_i$
   Unconstrained equation: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k + \epsilon_i$
2. Estimate the constrained equation and the unconstrained equation to compute:

$$F = \frac{(RSS_M - RSS)/M}{RSS/(N - K - 1)}$$

   where,
   M - degrees of freedom in numerator (nb. of constraints)
   $N - K - 1$ - degrees of freedom in the denominator
3. Reject H_0 if $F > F_c$
   Do not reject H_0 if $F < F_c$

F-test Steps

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$H_A : H_0 \text{ in not true}$$

1. Translate the null hypothesis into constraints on the equation. $\rightarrow$
   Constrained equation: $Y_i = \beta_0 + \epsilon_i$
   Unconstrained equation: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k + \epsilon_i$
2. Estimate the constrained equation and the unconstrained equation to compute:

   $$F = \frac{(RSS_M - RSS)/M}{RSS/(N - K - 1)}$$

   where,
   M - degrees of freedom in numerator (nb. of constraints)
   $N - K - 1$ - degrees of freedom in the denominator
3. Reject H_0 if $F > F_c$
   Do not reject H_0 if $F < F_c$

F-test Steps

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$H_A : H_0 \text{ in not true}$$

1. Translate the null hypothesis into constraints on the equation. $\rightarrow$
   Constrained equation: $Y_i = \beta_0 + \epsilon_i$
   Unconstrained equation: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k + \epsilon_i$
2. Estimate the constrained equation and the unconstrained equation to compute:

   $$F = \frac{(RSS_M - RSS)/M}{RSS/(N - K - 1)}$$

   where,
   M - degrees of freedom in numerator (nb. of constraints)
   $N - K - 1$ - degrees of freedom in the denominator
3. Reject H_0 if $F > F_c$
   Do not reject H_0 if $F < F_c$

F in the Regression Results

OLS Regression Results

Dep. Variable:	Profit	R-squared (uncentered):	0.984
Model:	OLS	Adj. R-squared (uncentered):	0.982
Method:	Least Squares	F-statistic:	839.8
Date:	Wed, 29 Sep 2021	Prob (F-statistic):	2.95e-49
Time:	11:27:08	Log-Likelihood:	-661.85
No. Observations:	60	AIC:	1332.
Df Residuals:	56	BIC:	1340.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
RD	0.7145	0.068	10.490	0.000	0.578	0.851
Admin	0.2546	0.043	5.927	0.000	0.169	0.341
Marketing	0.0962	0.025	3.915	0.000	0.047	0.145
Office	0.9425	0.663	1.421	0.161	-0.386	2.271

Omnibus:	6.429	Durbin-Watson:	1.745
Prob(Omnibus):	0.040	Jarque-Bera (JB):	9.643
Skew:	-0.196	Prob(JB):	0.00805
Kurtosis:	4.924	Cond. No.	91.1

OLS Assumptions

Notation	Meaning
1. $E(u_i) = 0$	Average value of residuals is zero
2. $Var(u_i) = \sigma^2$	The variance of the residuals is constant
3. $Cov(u_i, x_i) = 0$	There is no linear relationship between residues and x
4. $u_i \sim N(0, \sigma^2)$	Residuals have a normal distribution
5. $Cov(u_i, u_j) = 0$	Residuals don't depend on each other
6. Multicollinearity	Independent variables are not correlated with each other

Model Specification

Specifying an econometric equation consists of three parts:

1. Choosing the correct independent variables
2. Choosing the correct functional form
3. Choosing the correct form of the stochastic error term

Model Specification

Specifying an econometric equation consists of three parts:

1. Choosing the correct independent variables
2. Choosing the correct functional form
3. Choosing the correct form of the stochastic error term

Model Specification

Specifying an econometric equation consists of three parts:

1. Choosing the correct independent variables
2. Choosing the correct functional form
3. Choosing the correct form of the stochastic error term

Model Specification

Specifying an econometric equation consists of three parts:

1. Choosing the correct independent variables
2. Choosing the correct functional form
3. Choosing the correct form of the stochastic error term

Choosing the Correct Independent Variables

There are two mistakes one can make when choosing the independent variables for the model:

1. Omitting an important variable.
2. Including an irrelevant variable.

Choosing the Correct Independent Variables

There are two mistakes one can make when choosing the independent variables for the model:

1. Omitting an important variable.
2. Including an irrelevant variable.

Choosing the Correct Independent Variables

There are two mistakes one can make when choosing the independent variables for the model:

1. Omitting an important variable.
2. Including an irrelevant variable.

Omitting an Important Variable

Omitted Variable Bias

True Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

Specified Model

$$Y_i = \beta_0^* + \beta_1^* X_{1i} + \epsilon_i^*$$

$$\downarrow$$

$$\epsilon_i^* = \epsilon_i + \beta_2 X_{2i}$$

If X_1 and X_2 are correlated Cov $(X_{1i}, \epsilon^*) \neq 0 \rightarrow \underbrace{\hat{\beta}_1^* \neq \beta_1}_{\text{Biased!}}$

Omitting an Important Variable

Omitted Variable Bias

True Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

Specified Model

$$Y_i = \beta_0^* + \beta_1^* X_{1i} + \epsilon_i^*$$

$$\downarrow$$

$$\epsilon_i^* = \epsilon_i + \beta_2 X_{2i}$$

If X_1 and X_2 are correlated Cov $(X_{1i}, \epsilon^*) \neq 0 \rightarrow \underbrace{\hat{\beta}_1^* \neq \beta_1}_{\text{Biased!}}$

Omitting an Important Variable

Omitted Variable Bias

True Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

Specified Model

$$Y_i = \beta_0^* + \beta_1^* X_{1i} + \epsilon_i^*$$

$$\downarrow$$

$$\epsilon_i^* = \epsilon_i + \beta_2 X_{2i}$$

If X_1 and X_2 are correlated Cov $(X_{1i}, \epsilon^*) \neq 0 \rightarrow \underbrace{\hat{\beta}_1^* \neq \beta_1}_{\text{Biased!}}$

Including Irrelevant Variables

Including an irrelevant variable won't bias $\hat{\beta}_k$

True Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$$

Specified Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i^{**}$$

$$\downarrow$$

$$\epsilon_i^{**} = \epsilon_i - \beta_2 X_{2i}$$

$\hat{\beta}_1^* = \beta_1$ but Var $(\hat{\beta}_1) \uparrow$ thus $\underbrace{t_{X_1} \downarrow}_{\text{Biased!}}$

Also $R^2 \uparrow$, but $\bar{R}^2 \downarrow$

Including Irrelevant Variables

Including an irrelevant variable won't bias $\hat{\beta}_k$

True Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$$

Specified Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i^{**}$$

$$\downarrow$$

$$\epsilon_i^{**} = \epsilon_i - \beta_2 X_{2i}$$

$\hat{\beta}_1^* = \beta_1$ but Var $(\hat{\beta}_1) \uparrow$ thus $\underbrace{t_{X_1} \downarrow}_{\text{Biased!}}$

Also $R^2 \uparrow$, but $\bar{R}^2 \downarrow$

Including Irrelevant Variables

Including an irrelevant variable won't bias $\hat{\beta}_k$

True Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$$

Specified Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i^{**}$$

$$\downarrow$$

$$\epsilon_i^{**} = \epsilon_i - \beta_2 X_{2i}$$

$\hat{\beta}_1^* = \beta_1$ but Var $(\hat{\beta}_1) \uparrow$ thus $\underbrace{t_{X_1} \downarrow}_{\text{Biased!}}$

Also $R^2 \uparrow$, but $\bar{R}^2 \downarrow$

Including Irrelevant Variables

Including an irrelevant variable won't bias $\hat{\beta}_k$

True Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$$

Specified Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i^{**}$$

$$\downarrow$$

$$\epsilon_i^{**} = \epsilon_i - \beta_2 X_{2i}$$

$\hat{\beta}_1^* = \beta_1$ but Var $(\hat{\beta}_1) \uparrow$ thus $\underbrace{t_{X_1} \downarrow}_{\text{Biased!}}$

Also $R^2 \uparrow$, but $\bar{R}^2 \downarrow$

Including Irrelevant Variables

Including an irrelevant variable won't bias $\hat{\beta}_k$

True Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$$

Specified Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i^{**}$$

$$\downarrow$$

$$\epsilon_i^{**} = \epsilon_i - \beta_2 X_{2i}$$

$\hat{\beta}_1^* = \beta_1$ but Var $(\hat{\beta}_1) \uparrow$ thus $\underbrace{t_{X_1} \downarrow}_{\text{Biased!}}$

Also $R^2 \uparrow$, but $\bar{R}^2 \downarrow$

Specification Criteria

1. Theory: Is the variables place in the equation unambiguous and theoretically sound?
2. t-Test: Is the variables estimated coefficient significant in the expected direction?
3. $\bar{R}^2$: Does the overall fit of the equation (adjusted for degrees of freedom) improve when the variable is added to the equation?
4. Bias: Do other variables coefficients change significantly when the variable is added to the equation?

Specification Criteria

1. Theory: Is the variables place in the equation unambiguous and theoretically sound?
2. t-Test: Is the variables estimated coefficient significant in the expected direction?
3. $\bar{R}^2$: Does the overall fit of the equation (adjusted for degrees of freedom) improve when the variable is added to the equation?
4. Bias: Do other variables coefficients change significantly when the variable is added to the equation?

Specification Criteria

1. Theory: Is the variables place in the equation unambiguous and theoretically sound?
2. t-Test: Is the variables estimated coefficient significant in the expected direction?
3. $\bar{R}^2$: Does the overall fit of the equation (adjusted for degrees of freedom) improve when the variable is added to the equation?
4. Bias: Do other variables coefficients change significantly when the variable is added to the equation?

Specification Criteria

1. Theory: Is the variables place in the equation unambiguous and theoretically sound?
2. t-Test: Is the variables estimated coefficient significant in the expected direction?
3. $\bar{R}^2$: Does the overall fit of the equation (adjusted for degrees of freedom) improve when the variable is added to the equation?
4. Bias: Do other variables coefficients change significantly when the variable is added to the equation?

Choosing the Functional Form

Should we include the intercept β_0?

- Beginner researchers may want to supress the intercept when there is no logical interpretation for β_0.
- This would be a mistake. β_0 should be suppressed only on very rare occasions.

  For example: $C_i = \beta_0 + \beta_1 Q_i + \epsilon_i$

  where,

  C_i are the fixed and variable costs of businesses in a branch
- Excluding β_0 in this context would preserve one degree of freedom and would supposedly make the estimate of β_1 more accurate.
- However such examples in real life are rare!
- The risks associated with supressing β_0 most times outweight the benefits of doing so.

Choosing the Functional Form

Should we include the intercept β_0?

- Beginner researchers may want to supress the intercept when there is no logical interpretation for β_0.
- This would be a mistake. β_0 should be suppressed only on very rare occasions.
  For example: $C_i = \beta_0 + \beta_1 Q_i + \epsilon_i$
  where,
  C_i are the fixed and variable costs of businesses in a branch
- Excluding β_0 in this context would preserve one degree of freedom and would supposedly make the estimate of β_1 more accurate.
- However such examples in real life are rare!
- The risks associated with supressing β_0 most times outweight the benefits of doing so.

Choosing the Functional Form

Should we include the intercept β_0?

- Beginner researchers may want to supress the intercept when there is no logical interpretation for β_0.
- This would be a mistake. β_0 should be suppressed only on very rare occasions.

  For example: $C_i = \beta_0 + \beta_1 Q_i + \epsilon_i$
  where,
  C_i are the fixed and variable costs of businesses in a branch
- Excluding β_0 in this context would preserve one degree of freedom and would supposedly make the estimate of β_1 more accurate.
- However such examples in real life are rare!
- The risks associated with supressing β_0 most times outweight the benefits of doing so.

Choosing the Functional Form

Should we include the intercept β_0?

- Beginner researchers may want to supress the intercept when there is no logical interpretation for β_0.
- This would be a mistake. β_0 should be suppressed only on very rare occasions.

  For example: $C_i = \beta_0 + \beta_1 Q_i + \epsilon_i$

  where,

  C_i are the fixed and variable costs of businesses in a branch
- Excluding β_0 in this context would preserve one degree of freedom and would supposedly make the estimate of β_1 more accurate.
- However such examples in real life are rare!
- The risks associated with supressing β_0 most times outweight the benefits of doing so.

Choosing the Functional Form

Should we include the intercept β_0?

- Beginner researchers may want to supress the intercept when there is no logical interpretation for β_0.
- This would be a mistake. β_0 should be suppressed only on very rare occasions.

  For example: $C_i = \beta_0 + \beta_1 Q_i + \epsilon_i$

  where,

  C_i are the fixed and variable costs of businesses in a branch
- Excluding β_0 in this context would preserve one degree of freedom and would supposedly make the estimate of β_1 more accurate.
- However such examples in real life are rare!
- The risks associated with supressing β_0 most times outweight the benefits of doing so.

Choosing the Functional Form

Should we include the intercept β_0?

- Beginner researchers may want to supress the intercept when there is no logical interpretation for β_0.
- This would be a mistake. β_0 should be suppressed only on very rare occasions.

  For example: $C_i = \beta_0 + \beta_1 Q_i + \epsilon_i$

  where,

  C_i are the fixed and variable costs of businesses in a branch
- Excluding β_0 in this context would preserve one degree of freedom and would supposedly make the estimate of β_1 more accurate.
- However such examples in real life are rare!
- The risks associated with supressing β_0 most times outweight the benefits of doing so.

Choosing the Functional Form

Should we include the intercept β_0?

- Beginner researchers may want to supress the intercept when there is no logical interpretation for β_0.
- This would be a mistake. β_0 should be suppressed only on very rare occasions.

  For example: $C_i = \beta_0 + \beta_1 Q_i + \epsilon_i$

  where,

  C_i are the fixed and variable costs of businesses in a branch
- Excluding β_0 in this context would preserve one degree of freedom and would supposedly make the estimate of β_1 more accurate.
- However such examples in real life are rare!
- The risks associated with supressing β_0 most times outweight the benefits of doing so.

Choosing the Functional Form

Should we include the intercept β_0?

- Beginner researchers may want to supress the intercept when there is no logical interpretation for β_0.
- This would be a mistake. β_0 should be suppressed only on very rare occasions.

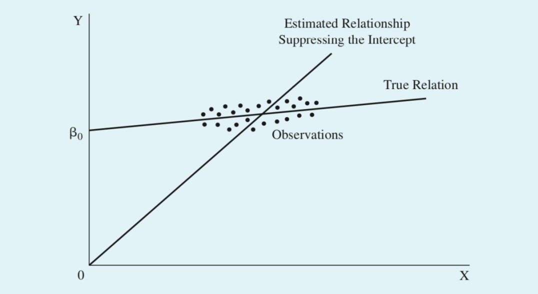
  For example: $C_i = \beta_0 + \beta_1 Q_i + \epsilon_i$

  where,

  C_i are the fixed and variable costs of businesses in a branch
- Excluding β_0 in this context would preserve one degree of freedom and would supposedly make the estimate of β_1 more accurate.
- However such examples in real life are rare!
- The risks associated with supressing β_0 most times outweight the benefits of doing so.

Suppressing β_0

By omitting the constant term, we force the impact of the constant into the estimates of the other coefficients.

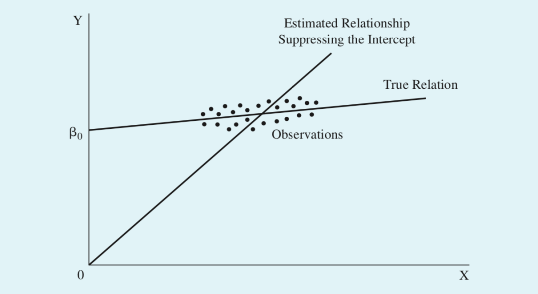


- The estimated regression will go through the origin
- $\hat{\beta}_1 \neq \beta_1$ - biassed coefficient
- t-scores $\uparrow$ - also biased

General Advice: Do Not Suppress the Intercept

Suppressing β_0

By omitting the constant term, we force the impact of the constant into the estimates of the other coefficients.

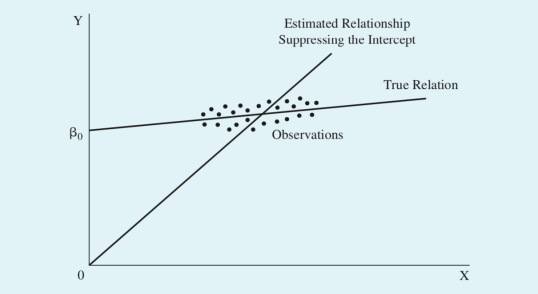


- The estimated regression will go through the origin
- $\hat{\beta}_1 \neq \beta_1$ - biassed coefficient
- t-scores ↑ - also biased

General Advice: Do Not Suppress the Intercept

Suppressing β_0

By omitting the constant term, we force the impact of the constant into the estimates of the other coefficients.

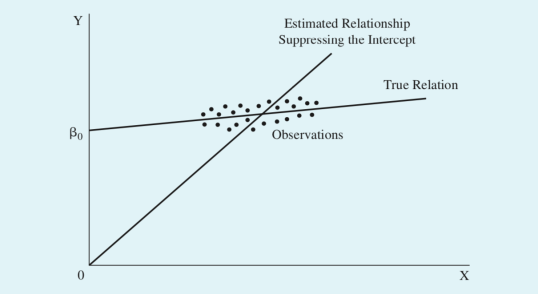


- The estimated regression will go through the origin
- $\hat{\beta}_1 \neq \beta_1$ - biassed coefficient
- t-scores $\uparrow$ - also biased

General Advice: Do Not Suppress the Intercept

Suppressing β_0

By omitting the constant term, we force the impact of the constant into the estimates of the other coefficients.



- The estimated regression will go through the origin
- $\hat{\beta}_1 \neq \beta_1$ - biassed coefficient
- t-scores $\uparrow$ - also biased

General Advice: Do Not Suppress the Intercept

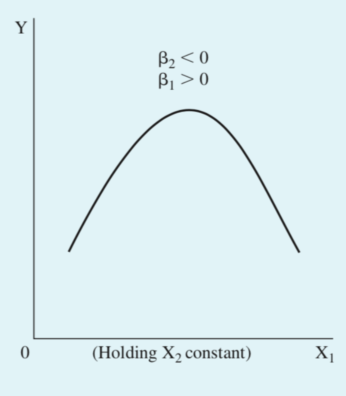
Polinomial Regressions

Sometimes the relationship between X and Y cannot be explained by a line.

Relationship between Earnings and Age.

- As a young worker gets older, his or her earnings will increase.
- Beyond some point an increase in age will not increase earnings, and around retirement earnings will start to fall abruptly with age.

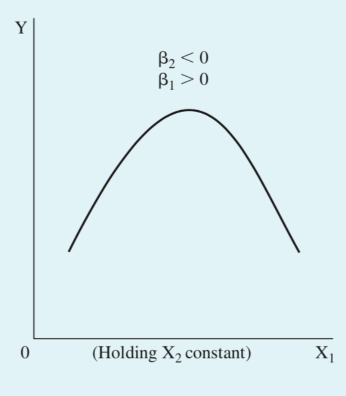
Relationship between Earnings and Age



$$\text{Earnings}_i = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Age}_i^2 + \cdots + \epsilon_i \qquad \frac{\Delta Y}{\Delta X_1} = \beta_1 + 2\beta_2 X_1$$

General Advice: Rely of Theory when Defining the Functional Form of Your Model.

Relationship between Earnings and Age



$$\text{Earnings}_i = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Age}_i^2 + \cdots + \epsilon_i \qquad \frac{\Delta Y}{\Delta X_1} = \beta_1 + 2\beta_2 X_1$$

General Advice: Rely of Theory when Defining the Functional Form of Your Model.

Choosing the correct form of the stochastic error term

1. Multicollinearity
2. Serial Correlation $Cov(u_i, u_j) \neq 0$
3. Heteroskedasticity $Var(u_i) \neq \sigma^2$

Choosing the correct form of the stochastic error term

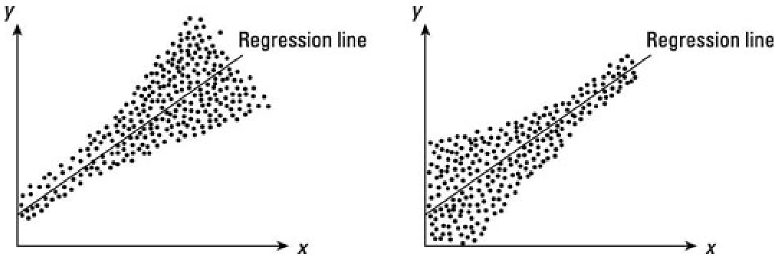
1. Multicollinearity
2. Serial Correlation $Cov(u_i, u_j) \neq 0$
3. Heteroskedasticity $Var(u_i) \neq \sigma^2$

Choosing the correct form of the stochastic error term

1. Multicollinearity
2. Serial Correlation $Cov(u_i, u_j) \neq 0$
3. Heteroskedasticity $Var(u_i) \neq \sigma^2$

Choosing the correct form of the stochastic error term

1. Multicollinearity
2. Serial Correlation $Cov(u_i, u_j) \neq 0$
3. Heteroskedasticity $Var(u_i) \neq \sigma^2$



Let's get Started!

Access Google Colaboratory through your Gmail account